DreamFuse: Adaptive Image Fusion with Diffusion Transformer

Junjia Huang[1,2*] Pengxiang Yan[3*] Jiyang Liu[3*†] Jie Wu[3]
Zhao Wang[3] Yitong Wang[3] Liang Lin[1,2,4] Guanbin Li[1,2,4‡]
[1]Sun Yat-sen University, [2]Peng Cheng Laboratory, [3]ByteDance Intelligent Creation
[4]Guangdong Key Laboratory of Big Data Analysis and Processing
huangjj77@mail2.sysu.edu.cn, wantong1017@163.com
linliang@ieee.org, liguanbin@mail.sysu.edu.cn
{yanpengxiang.ai, liujiyang.liu, wujie.10, zhaoxu.bit}@bytedance.com
https://ll3rd.github.io/DreamFuse/

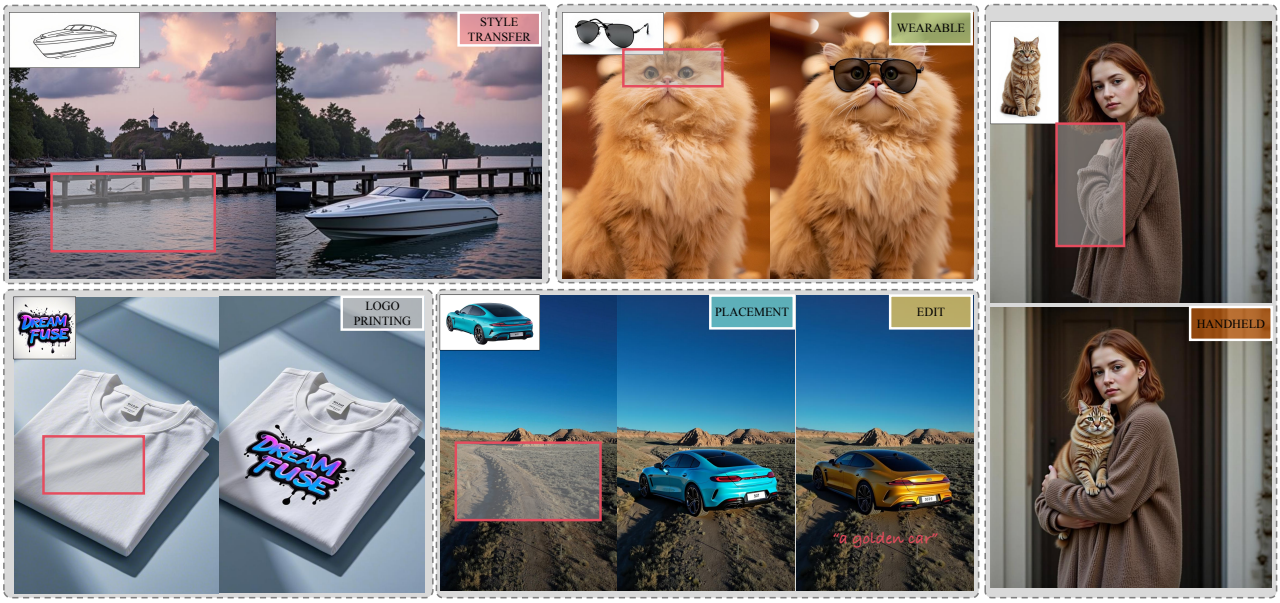

Figure 1. DreamFuse demonstrates adaptive performance across diverse scenarios, including style transfer, wearable items, logo printing, placement and handheld. Notably, when given a text prompt, our method effectively responds by further editing the attributes of the foreground object (e.g., a golden car).

Abstract

Image fusion seeks to seamlessly integrate foreground objects with background scenes, producing realistic and harmonious fused images. Unlike existing methods that directly insert objects into the background, adaptive and interactive fusion remains a challenging yet appealing task. It requires the foreground to adjust or interact with the background context, enabling more coherent integration. To address this, we propose an iterative human-in-the-loop data generation pipeline, which leverages limited initial data with diverse textual prompts to generate fusion datasets across various scenarios and interactions, including placement, holding, wearing, and style transfer. Building on this, we introduce DreamFuse, a novel approach based on the Diffusion Transformer (DiT) model, to generate consistent and harmonious fused images with both foreground and background information. DreamFuse employs a Positional Affine mechanism to inject the size and position of the foreground into the background, enabling effective foreground-background interaction through shared attention. Furthermore, we apply Localized Direct Preference Optimization guided by human feedback to refine DreamFuse, enhancing background consistency and foreground harmony. DreamFuse achieves harmonious fusion while generalizing to text-

*Equal Contribution. †Project Lead. ‡Corresponding Author.

driven attribute editing of the fused results. Experimental results demonstrate that our method outperforms state-of-the-art approaches across multiple metrics.

1. Introduction

Foreground-background image fusion is a fundamental and practical task in image editing. Recently, with the rapid development of generation methods based on diffusion models, numerous innovative and imaginative approaches have emerged in this field. Beyond generating images purely driven by text prompts, increasing attention has been directed toward generating customized images guided by specific foreground objects [1, 28, 32, 53] or performing inpainting on designated background regions [45] for image editing and fusion. Going further, several methods aim to achieve harmonious fusion of foreground and background by adjusting details such as lighting [50], shadows [34], and brightness-contrast consistency [38], making the fused images appear more natural. Other approaches [2, 7, 42] strive to enhance fusion by modifying the orientation, pose, or style of the foreground object while preserving its identity attributes, enabling better adaptation to the background. However, most of these methods typically focus on directly placing the foreground object into the background scene. In contrast, practical scenarios often involve more diverse and interactive cases, such as partial occlusions, alternating visibility, or interactions where the object is held, worn, or integrated into the scene.

A major challenge in handling such complex fusion scenarios is the lack of suitable datasets. Existing methods typically rely on object segmentation from images or videos [13, 22] followed by inpainting [27, 31] to reconstruct backgrounds. However, this multi-step process frequently suffers from quality degradation due to imprecise segmentation or suboptimal inpainting, which can introduce artifacts like shadows or residual elements in the background. Moreover, handling partially occluded foregrounds is particularly challenging [36], and segmentation-based data often fails to adjust object pose or perspective to align with the background during fusion. Based on these observations, we propose an Iterative Human-in-the-Loop Data Generation Pipeline to directly generate fused data, including foregrounds, backgrounds, and the fused images, avoiding issues such as incomplete foregrounds and background artifacts. We train a DiT model on curated fused data with text prompts, modifying its attention mechanism to shared attention to ensure identity consistency across fused data. Using this model, we generate multi-scale fused data for various scenarios with diverse prompts, such as placement, handheld interactions, wearable items, logo printing, and style transfer. Throughout the process, we enhance content diversity by incorporating existing LoRA [10] and iteratively optimize the model through manual selected data. We utilize GPT-4o to filter out low-quality fused data, such as mismatched foregrounds or degraded images, ultimately constructing a dataset of 80,000 high-quality, multi-scene, multi-scale fused image samples.

Another critical challenge in image fusion is ensuring background consistency and foreground harmony. Some approaches [2, 7] rely on masks or bounding boxes for foreground placement, blending the background outside the mask to maintain consistency. However, these approaches often fail to realistically render effects like shadows or reflections beyond the mask, limiting the realism of the fused results. Other methods [20, 40] reconstruct fused images through inversion, improving foreground harmony but often compromising background consistency. To address this trade-off, we propose DreamFuse, an adaptive image fusion framework based on DiT. By incorporating shared attention, we condition the the fused image generation on both the foreground and background while employing positional affine to introduce the foreground's position and scale without restricting its editable regions. Additionally, we employ Localized Direct Preference Optimization (LDPO) to further optimize the foreground and background regions of the fused image, ensuring better alignment with human preferences. Experimental results demonstrate that DreamFuse performs exceptionally well across various scenarios. As shown in Fig. 1, DreamFuse produces highly realistic fusion effects for real-world images. Furthermore, during training, a certain proportion of fused image descriptions is incorporated as prompts. When given a text prompt, DreamFuse effectively responds to the input and enables attribute modifications in the fused scenes, such as turning a car into gold.

In summary, our key contributions are threefold:

- We propose an iterative Human-in-the-Loop data generation pipeline and construct a comprehensive fusion dataset containing 80k diverse fusion scenarios.
- We propose DreamFuse, a fusion framework based on DiT, which leverages positional affine and LDPO strategies to integrate the foreground into the background more naturally and adaptively.
- Our method outperforms the state-of-the-art methods on various benchmarks and remains effective in real-world and out-of-distribution scenarios.

2. Related Work

Customized Image Generation. Customized image generation aims to create user-specific images based on text prompts or reference images. Some methods [5, 6, 14, 28] incorporate reference concepts into specific text prompts, while approaches [8, 39, 46, 52] utilize additional encoders to encode reference images as visual prompts, introducing customized representations for generation. Other methods [12, 32] achieve subject-driven generation by directly

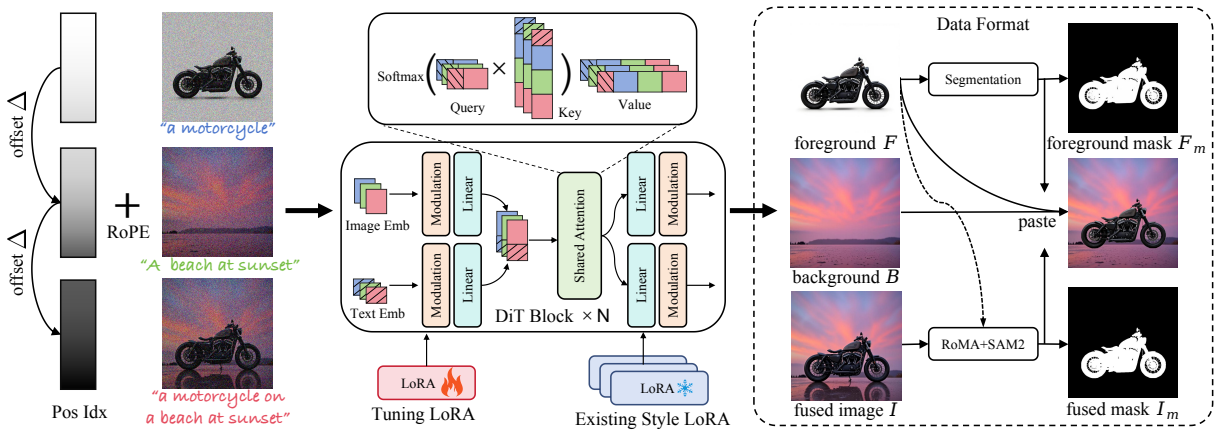

Figure 2. The framework of the data generation model and position matching process. The left side of the image illustrates the design structure of our data generation model, while the right side shows the position matching process and data format. We enhance the diversity of fused data generation through flexible and rich prompts combined with various style LoRAs.

concatenating reference and target images during generation. In this paper, we further extend generative diffusion models by fine-tuning on small-scale data to directly generate customized foregrounds, backgrounds, and fused images based on corresponding text prompts.

Image Fusion. The goal of image fusion is to seamlessly integrate an object from a foreground image into a background image. Compared to directly cutting and pasting, some approaches [21, 24, 50, 54] adjust the lighting, shadows, and colors of the pasted foreground region to achieve a more harmonious fusion. Other methods [2, 7, 20, 33, 40–42] focus on altering the perspective, pose, or style of the foreground object to make it fit more naturally into the background image. However, these methods are often restricted to object placement. In this paper, we propose a more versatile fusion approach that supports a variety of scenarios, including hand-held objects, wearable items, and style transformations, enabling more diverse integrations.

Human Feedback Learning. Many methods now leverage human feedback learning to make generated images more aligned with users' preferences. Some approaches [18, 43, 44, 48, 49] train a reward model to understand human preferences and improve the generation quality through reward feedback learning. Others [16, 37, 51] utilize human comparison data to directly optimize a policy that best satisfies human preferences. For the image fusion task, we propose localized direct preference optimization, which focuses on region-specific optimization to enhance both the background consistency and the harmony in fused images.

3. Methodology

As shown in the data format in Fig. 2, the image fusion task typically involves the following types of images: a foreground image $F \in \mathbb{R}^{H\times W\times 3}$ with mask $F_m \in \mathbb{R}^{H\times W}$, a background image $B \in \mathbb{R}^{H\times W\times 3}$, a fused image $I \in \mathbb{R}^{H\times W\times 3}$, a fused mask $I_m \in \mathbb{R}^{H\times W}$ associated with the fused object. The masks F_m and I_m are primarily used to indicate the position and size of the foreground object. In practical applications, only the centroid and bounding box of the mask are required.

3.1. Iterative Human-in-the-Loop Data Generation

Data Startup. Unlike methods [7, 35] that start with a fused image to segment the foreground and generate the background using inpainting, we aim to create higher-quality fusion data with richer scenes and more diverse foreground fusion. To this end, we design an iterative, human-in-the-loop data generation process. We first extract a pair of high-quality foreground F and fused image I from a subject-driven dataset [32]. We then manually refine the inpainting regions to remove the foreground object and its effects, such as reflections and shadows, creating a high-quality background image B. A total of 86 initial samples are curated, and their corresponding descriptions C are generated with GPT-4o. These data are then fed into the data generation model depicted in Fig. 2 for training.

Data Generation Model Design. We adopt Flux [15] as the base model and input our curated fusion samples $G = (F, B, I)$ with prompts $C_G = (C_F, C_B, C_I)$ in batches. The images and prompts are encoded into image embeddings $E_i \in \mathbb{R}^{h\times w\times d}$ and text embeddings E_c, supplemented by learnable tag embeddings to differentiate the foreground and background. In Flux's RoPE [30] mechanism, which uses a 2D position index $P_{idx} = (i, j), \forall i \in [0, h), j \in [0, w)$ to represent image positions, we optionally introduce an offset Δ to P_{idx} for F, B, and I as follows:

$$P_{idx} = \begin{cases} (i,j), & \text{if } F, \\ (i,j)+\Delta, & \text{if } B, \\ (i,j)+2\Delta, & \text{if } I. \end{cases} \tag{1}$$

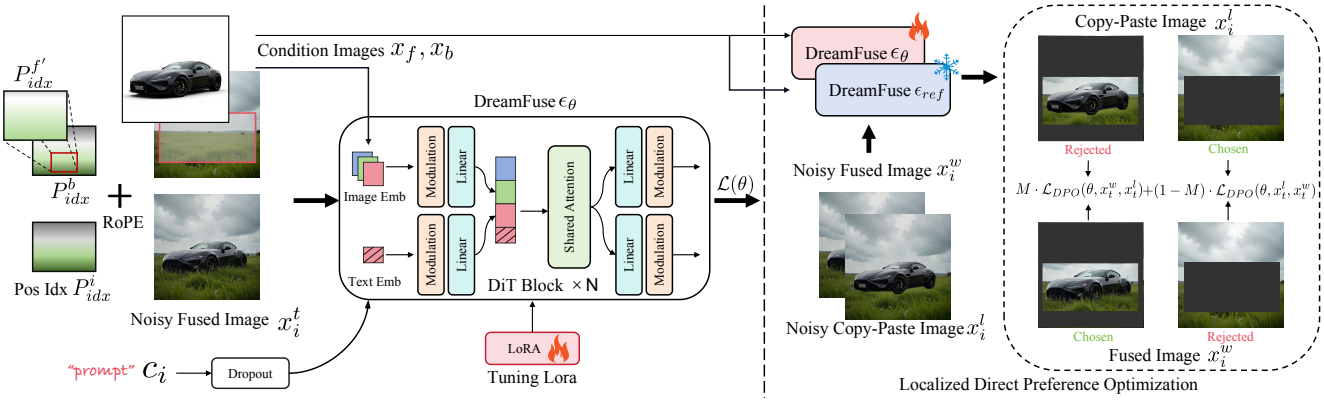

Figure 3. The framework of the DreamFuse. We apply positional affine transformations to map the foreground's position and size onto the background. The foreground and background are concatenated with the noisy fused image as condition images before DiT's attention layers. Localized direct preference optimization is then used to improve background consistency and foreground harmony.

During training, adding an offset improves the model's ability to generate diverse fused scenes but performs poorly across resolutions, while omitting the offset produces overly consistent results yet adapts well to multi-scale data. Based on this, we use two models—with and without offset—to generate diverse, multi-scale samples. Further details are provided in the supplementary materials.

To establish connections between F,B and I, we modify the original independent attention mechanism in the DiT to a shared attention (SA) mechanism [11]. As shown in Fig. 2, after the text embeddings and image embeddings of each sample are processed through modulation and linear layers, they are concatenated to form the attention query $Q_g = [Q_g^c; Q_g^i], g \in G$, where c and i represent the text and image components. We then concatenate the image components of the key and value across all samples: $K_g = [K_g^c; K_F^i; K_B^i; K_I^i]$, $V_g = [V_g^c; V_F^i; V_B^i; V_I^i]$. The shared attention is then computed as:

$$SA = softmax(\frac{Q_g K_g^T}{\sqrt{d}})V_g, \tag{2}$$

where d denotes the dimension. After applying shared attention, the model gains an initial ability to generate similar images. We then fine-tune the model with LoRA [10] to enable the generation of high-quality image fusion samples.

Scene Generalizability and Style Variability. Our initial data only consists of object placement scenes. After fine-tuning the data generation model, it demonstrates the ability to generalize to more diverse scene prompts. In subsequent iterations, we leverage GPT-4o with open-source prompts to generate fusion prompts C_G, expanding foreground objects to include animals, pets, products, portraits, and logos, while categorizing background scenes into indoor and outdoor settings. Fusion scenarios are further diversified to include placement, handheld, logo printing, wearable, and style transfer. Furthermore, we observe that the data generation model responds effectively to existing style LoRAs. Therefore, we integrate fine-tuned style LoRAs, such as depth of field, realism, and ethnicity, to further enhance fusion data in both scene variety and artistic style, mitigating the stylistic bias of the Flux base model. This expansion process is iterative: the model is fine-tuned on data generated in the previous step, additional data is generated, and high-quality fusion samples are manually curated to serve as input for the next round of fine-tuning.

Position Matching. To determine the position of the foreground object in the background, we use RoMA [4] to perform feature matching between the foreground and the fused image, converting the results into bounding boxes. Then we utilize SAM2 [26] to segment the foreground object from the fused image based on these bounding boxes, yielding I_m. For the foreground object, we use an internal segmentation model to obtain F_m. The position and size of the foreground object in the background are then calculated using the centroids and bounding boxes of I_m and F_m.

3.2. Adaptive Image Fusion Framework

In image fusion tasks, three key aspects need to be considered: (1) how to model the relationship between the background, foreground, and fused image; (2) how to incorporate the position and size information of the foreground object into the background; (3) how to ensure background consistency and foreground harmony in the fused image.

Condition-aware Modeling. Inspired by the work [32], we model the background and foreground as conditions, with the fused image treated as the denoised target. Given a fixed dataset $\mathcal{D} = \{(c_i, x_f, x_b, x_i)\}$, each sample consists of a textual description of the fused image c_i, along with images representing the foreground x_f, background x_b and fused image x_i. We adopt the Flux-based DiT architecture, using the foreground x_f and background x_b as conditions with a fix timestep 0. Additionally, most text prompt c_i are

randomly dropped out with a probability p during training, replaced with empty strings, while a portion of the prompts is retained to preserve the network's text-responsive capability. The DiT network is tasked with denoising the noisy fused image x_i^t at timestep t defined as:

$$x_i^t = (1-t)x_i + tx_n, \tag{3}$$

where $x_n \sim q(x_n)$ denotes a noise sample and $t \in [0,1]$. The DiT model is trained to regress the velocity field $\epsilon_\theta(x_i^t, x_f, x_b, t)$ by minimizing the Flow Matching [17] objective $\mathcal{L}_{noise}(\theta)$:

$$\mathbb{E}_{t,(c_i,x_f,x_b,x_i)\sim\mathcal{D},x_n\sim q(x_n)}[||\epsilon - \epsilon_\theta(c_i, x_i^t, x_f, x_b, t)||], \tag{4}$$

where the target velocity field is $\epsilon = x_n - x_i$. Within the DiT attention mechanism, all components are concatenated as $[Dropout(c_i, p), x_i, x_f, x_b]$, enabling joint attention computation, as illustrated in Fig. 3. This mechanism effectively integrates background and foreground information into the fused image through the attention layers.

Positional Affine. We explore three approaches to incorporate positional information, as shown in Fig. 4. The most straightforward approach is to directly transform the foreground to match the desired position and size in the background (Fig. 4 (b)). However, this method compresses the information of foreground during scaling, which is unfavorable for inserting small objects. Another approach involves using the placement information of the foreground, such as the mask after positioning, as a condition. This information is encoded via a tokenizer and introduced into the attention computation (Fig. 4 (c)). However, this approach relies heavily on the tokenizer, requiring a large amount of data to optimize its representation of positional information. To leverage the relative positional relationship of the foreground more directly and effectively, we propose the positional affine method shown in Fig. 4 (a).

Specifically, both the foreground and background are assigned 2D position index $P_{idx}^f, P_{idx}^b = (i,j), \forall i \in [0,h), j \in [0,w)$ to represent their spatial relationships within the image. When placing the foreground in a target region $P_{idx}^r = (u,v), \forall u \in [h_r, h_r'), v \in [w_r, w_r')$ of the background, the affine transformation matrix A is computed as follows:

$$A = \begin{bmatrix} \frac{w_r'-w_r}{w} & 0 & w_r \\ 0 & \frac{h_r'-h_r}{h} & h_r \\ 0 & 0 & 1 \end{bmatrix}. \tag{5}$$

Next, the position index P_{idx}^r of the target region is mapped to the foreground with the inverse affine transformation:

$$P_{idx}^{f'} = A^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}. \tag{6}$$

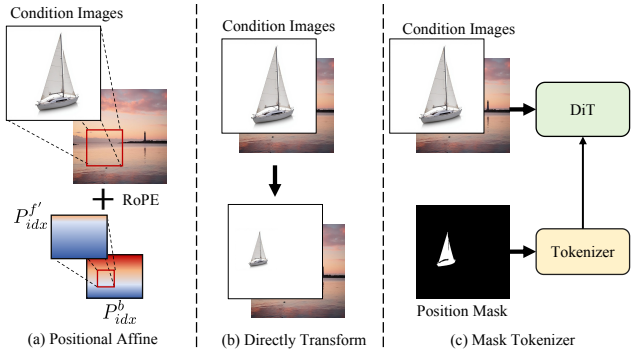

Figure 4. Three ways for injecting positional conditions: (a) using positional affine to map the foreground's position index to its target placement; (b) directly transforming the foreground object to the target position; (c) encoding position mask information with a tokenizer and integrating it into DiT's attention computation.

We utilize $P_{idx}^{f'}$ as the new position index of foreground. By employing this positional affine transformation, and leveraging DiT's responsiveness to position index, we directly incorporate the position and size information of the foreground into the target location within the background. This approach eliminates the need to scale or compress the foreground, enabling a more effective and reasonable integration of positional information.

Localized Preference Optimization. In the image fusion process, maintaining background consistency and foreground harmony is crucial. When directly generating the denoised fused image, issues such as inconsistent backgrounds or disharmonious foregrounds can easily arise. To address this, we propose Localized Direct Preference Optimization (LDPO) based on Diffusion-DPO [18, 37], enabling the diffusion network to more effectively learn from human preferences in the context of image fusion.

We construct a dataset $\mathcal{D}' = \{(c_i, x_f, x_b, x_i^w, x_i^l)\}$ consisting of additional fused sample pairs, where x_i^w aligns better with human preferences than x_i^l. For example, as shown in Fig. 3, x_i^l can be a fused image obtained by directly copying and pasting the foreground onto the background. For simplicity, we define the model input as $x_t^{\{w,l\}} = (c_i, x_f, x_b, x_i^{t,\{w,l\}})$. The Diffusion-DPO optimizes a policy to satisfy human preferences via objective $\mathcal{L}_{DPO}(\theta, x_t^w, x_t^l)$:

$$\mathbb{E}\Big[\log\sigma\Big(-\frac{\beta}{2}(||\epsilon^w - \epsilon_\theta(x_t^w,t)||^2 - ||\epsilon^w - \epsilon_{ref}(x_t^w,t)||^2 - (||\epsilon^l - \epsilon_\theta(x_t^l,t)||^2 - ||\epsilon^l - \epsilon_{ref}(x_t^l,t)||^2))\Big)\Big], \tag{7}$$

where $\epsilon_\theta(\cdot)$ and $\epsilon_{ref}(\cdot)$ denotes the predictions of optimized model and reference model, respectively, β is the regularization coefficient and σ is the sigmoid function. Intuitively, minimizing $\mathcal{L}_{DPO}$ encourages the predicted velocity field

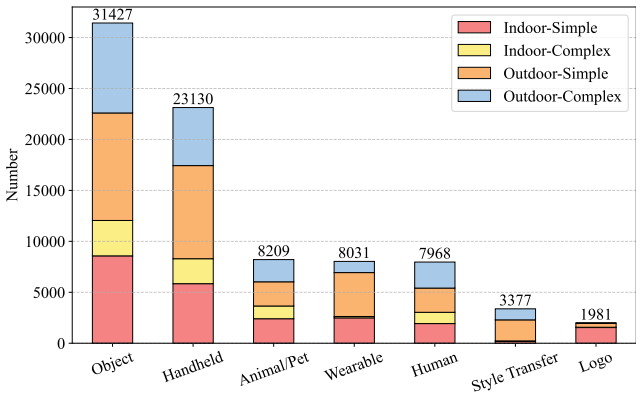

Figure 5. Scene distribution of the fusion dataset, including scenario counts, indoor/outdoor background proportions, and complexity levels.

ϵ_θ closer to the target velocity ϵ^w of the chosen data, while diverging from ϵ^l (the rejected data). However, not all aspects of x_i^l fail to align with human preferences. For instance, in copy-paste fused images, the consistency of the background better aligns with human preferences. Therefore, we adopt a Localized DPO strategy for x_i^w and x_i^l. A localized foreground region $M(f)$ is defined as:

$$M(f) = \begin{cases} 1, \text{if} f \in \alpha \cdot Bbox(x_f) \\ 0, \text{otherwise}, \end{cases} \tag{8}$$

where f represents a pixel location, $Bbox(x_f)$ denotes the region in the bounding box of the foreground object and α is a dilation factor that moderately expands this region. The optimized objective $\mathcal{L}_{LDPO}(\theta, x_t^w, x_t^l, M)$ is defined as:

$$M \cdot \mathcal{L}_{DPO}(\theta, x_t^w, x_t^l) + (1 - M) \cdot \mathcal{L}_{DPO}(\theta, x_t^l, x_t^w). \tag{9}$$

We provide the pseudo-code for LDPO in the Appendix. This strategy ensures background consistency while making the foreground more harmonious and aligned with human preferences.

4. Experiments

4.1. Dataset Analysis

As outlined in Sec. 3.1, we employ an iterative human-in-the-loop data generation pipeline to create approximately 400k image fusion samples. To ensure diversity, we leverage GPT-4o and existing prompt libraries during the prompt generation process to create a wide variety of prompts. After applying quality filtering techniques, including GPT-4o screening and gradient comparison, we curate a final dataset containing 80k high-quality fused image samples. We further analyze the fusion scenarios, background types, and complexity of the filtered dataset, as illustrated in Fig. 5. Notably, over half of the dataset features outdoor backgrounds, and approximately 23k images include hand-held

Method	CLIP	AES	IR	VR
ControlCom [47]	31.28	6.09	-0.27	1.33
TF-ICON [20]	**32.28**	<u>6.51</u>	<u>0.266</u>	<u>2.37</u>
Anydoor [2]	31.80	6.27	0.02	1.31
MADD [7]	30.36	5.94	-0.72	0.61
MimicBrush [3]	31.71	6.27	-0.33	0.83
Ours	<u>31.93</u>	**6.55**	**0.35**	**3.27**

Table 1. Quantitative evaluation results on TF-ICON dataset.

Method	CLIP	AES	IR	VR
ControlCom [47]	34.23	5.76	0.48	2.20
TF-ICON [20]	<u>35.21</u>	<u>6.06</u>	0.86	3.88
Anydoor [2]	35.04	6.01	<u>1.01</u>	<u>4.14</u>
MADD [7]	33.49	5.66	0.28	-0.08
MimicBrush [3]	34.46	6.01	0.56	3.19
Ours	**35.52**	**6.10**	**1.19**	**5.56**

Table 2. Quantitative evaluation results on DreamFuse test dataset.

scenarios. Details of the generation pipeline, quality filtering methods, and statistical results are provided in the supplementary materials.

4.2. Implementation Details

Hyperparameters. In DreamFuse training, we adopt Flux-Dev [15] as the base model. The input images are scaled proportionally from their original resolution to a short side of 512 pixels. The training procedure comprises two stages: first, the full DreamFuse 80k dataset is trained for 10k iterations with batch size 1, the Prodigy [23] optimizer, and a LoRA rank of 16. Subsequently, we manually select 15k high-quality samples from the initial dataset for LDPO training. Negative samples x_i^l in LDPO training consist of two types: copy-paste images and low-quality inference results generated from the selected samples after stage one. We apply the $\mathcal{L}_{LDPO}$ loss to the copy-paste data and $\mathcal{L}_{DPO}$ loss to remaining negative samples. The second stage employs the AdamW [19] optimizer (learning rate 5×10^{-5}) for another 10k iterations. The dropout probability p and dilation factor α is set to 99% and 1.5. The entire training is conducted on 8 NVIDIA A100 GPUs, requiring approximately 24 hours.

Benchmarks. We randomly selected 500 unseen samples from the generated dataset as DreamFuse test set to evaluate the model's fusion capabilities across multiple scenarios, including object placement, wearing, logo printing, handheld and style transfer. Additionally, we evaluated method's performance on out-of-domain data with the TF-ICON [20] dataset, which consists of 332 samples spanning four visual domains: photorealism, pencil sketching, oil painting, and cartoon animation.

Evaluation metrics. We utilize the CLIP [25] score to evaluate the similarity between the fused images and their corresponding descriptive texts, the AES [29] score to assess the aesthetic quality of the fused images, and the Im-

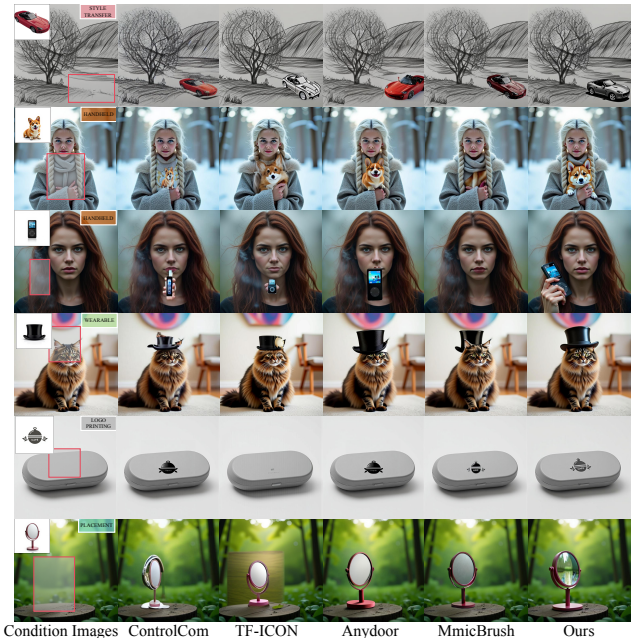

Figure 6. Qualitative comparisons with existing methods. The first row is from the TF-ICON dataset, while the others are from the DreamFuse test set. Our approach achieves a more seamless integration of foreground objects with the background images, resulting in higher visual consistency and realism.

ageReward [43] (IR) score to evaluate alignment, fidelity, and harmlessness. Additionally, we employ the VisionReward [44] (VR) score, which leverages a vision language model [9] (VLM) to evaluate the fused results from multiple perspectives across various questions, better reflecting human preferences.

4.3. Comparisons with Existing Methods

Quantitative results. We evaluate several existing state-of-the-art image fusion methods on the TF-ICON and DreamFuse test datasets, including image composition methods such as TF-ICON [20], Anydoor [2], MADD [7] and ControlCom [47], as well as the reference-based fusion method MimicBrush [3] . As shown in Tab. 1, our method outperforms existing approaches across multiple metrics on the TF-ICON dataset, with a notable 0.9 improvement in the VR score compared to the second-best method. The TF-ICON contains both realistic images and images from diverse domain, and our results highlight the robustness and generalization capability of our approach. Similarly, as shown in Tab. 2, our method achieves superior performance on the DreamFuse test set, surpassing the second-best method by 1.4. Our method consistently demonstrates better fusion quality across all scenarios.

Qualitative results. Fig. 6 presents the qualitative visualization results of our method compared with other methods. As shown, our method not only performs well across various fusion scenarios but also excels when the fused foreground object contains reflective surfaces. Specifically, for elements like the mirror in the last row, DreamFuse can perceive the surrounding environment and adaptively adjust the reflections on the mirror, resulting in more natural fusion image. Detailed discussions on fusion effects in real-world scenarios are provided in the supplementary materials.

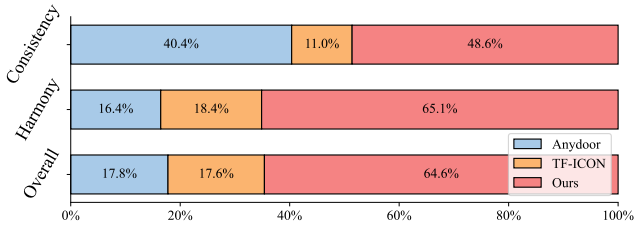

Figure 7. User Study: Evaluation from three perspectives: consistency, harmony, and overall quality.

Method	CLIP	AES	IR	VR
Direct Trans.	35.04	6.10	0.98	4.72
Mask Token.	35.14	6.11	1.03	4.93
Pos. Affine	**35.22**	**6.13**	**1.07**	**5.28**

Table 3. Qualitative analysis of three positional incorporation strategies on the DreamFuse test set after the first-stage training.

DreamFuse	CLIP	AES	IR	VR
First-Stage Training	35.22	**6.13**	1.07	5.28
w/ $\mathcal{L}_{DPO}$	35.36	6.07	1.16	5.52
w/ $\mathcal{L}_{DPO} + \mathcal{L}_{LDPO}$	35.51	6.10	1.17	**5.58**
w/ $\mathcal{L}_{DPO} + \mathcal{L}_{LDPO} + \mathcal{L}_{noise}$	**35.52**	6.10	**1.19**	5.56

Table 4. Qualitative analysis of localized preference optimization.

User study. As shown in Fig. 7, we randomly select 200 samples and conduct a user study with 17 participants to compare our method with Anydoor [2] and TF-ICON [20]. The evaluation focus on three perspectives: the consistency of the fused image's foreground and background with the input, the harmony of the fusion image, and the overall fusion quality. The results demonstrate that our method outperforms existing approaches across all three dimensions, achieving a 64.6% score in overall fusion quality.

4.4. Ablation Study

Positional affine. We compare the three positional incorporation strategies discussed in Sec. 3.2. As shown in Tab. 3, the results on the DreamFuse test set after the first-stage training demonstrate that the positional affine approach achieves the highest CLIP and VR scores, reaching 35.215 and 5.278, respectively. This suggests that the positional affine method better preserves the information of foreground objects, outperforming the direct transform and mask tokenizer strategies, which are more prone to information loss of fine-grained details.

Localized preference optimization. In the second stage of

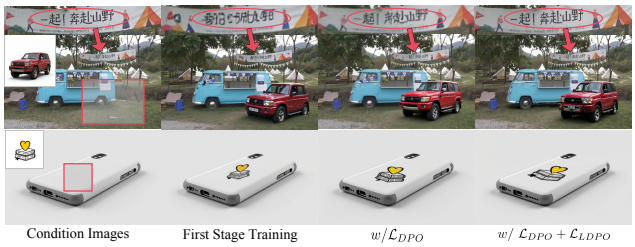

Figure 8. Qualitative results about the LDPO. Compared to "$w/\ \mathcal{L}_{DPO}$", "$w/\ \mathcal{L}_{LDPO}$" leverages copy-paste data to better help the model understand perspective changes in the foreground while maintaining background consistency as much as possible.

training, we employ Localized Direct Preference Optimization (LDPO) to further enhance the model's performance by utilizing two types of paired data: the first type consists of poorly performing and well-performing results generated after the first training stage with different seeds and optimized with the $\mathcal{L}_{DPO}$, while the second type involves copy-paste data treated as negative samples and optimized with the $\mathcal{L}_{LDPO}$. As shown in Tab. 4, "$w/\ \mathcal{L}_{DPO}$" refers to training with only the first type of data, "$w/\ \mathcal{L}_{LDPO}$" refers to training with copy-pasted data. Results indicate the LDPO significantly improves fusion performance by achieving an approximate 0.11 increase in IR scores compared to the first-stage results. Further incorporating the $\mathcal{L}noise$ loss results in a comparable performance. LDPO primarily enhances background consistency and foreground harmony, as visually demonstrated in Fig. 8, where using copy-paste data as negative samples allows the model to better capture fusion-related transformations, such as perspective and affine adjustments, rather than simply copying and pasting the foreground, while also improving the consistency of other background regions.

Dropout probability p**.** To retain the diffusion model's text response capability, allowing it to edit the fused foreground based on the prompt, we include the fusion image's text as a prompt at a dropout probability p. However, we find that when the dropout probability is set too low, such as 80%, the model's response to empty text weakens, resulting in the inability to properly integrate the foreground into the background. As shown in Fig. 9 (a), lower dropout probabilities lead to lower CLIP scores for the fusion results, indicating that the foreground is not well integrated into the background. Therefore, we ultimately chose $p = 99\%$, which preserves the model's text response capability to a greater extent without compromising its performance.

Dilation factor α**.** Intuitively, the dilation factor α defines the size of the local harmonious region around the foreground considered during the LDPO process. As shown in Fig. 9 (b), we investigate the impact of α on the model's performance. When $\alpha = 1$, the model focuses solely on the

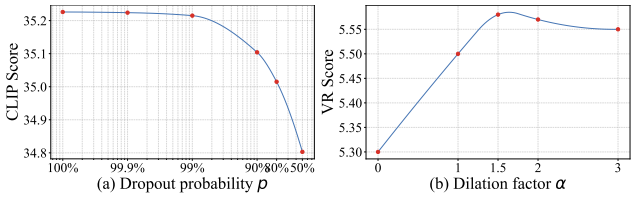

Figure 9. The effectiveness about the dropout probability p and dilation factor α.

Figure 10. The responsiveness to different prompts.

harmony within the bounding box of the object, disregarding external factors such as shadows or reflections, which are then treated as non-preferred, leading to a performance drop. When $\alpha = 0$, copy-paste data is entirely treated as samples consistent with human preferences, resulting in a significant decline in the VR score. As α increases, the preference for background consistency gradually diminishes, though it has minimal impact on overall harmony. Based on these findings, we set $\alpha = 1.5$ for LDPO training.

Responsiveness to text prompts. In our experiments, we find that DreamFuse responds effectively to text prompts without additional training. This enables modifications to the attributes of foreground objects or the fusion scene, as shown in Fig. 10.

4.5. Conclusion

In this paper, we first propose an iterative human-in-the-loop data generation pipeline to create fusion scenarios that are relatively rare in traditional fusion tasks, making the fusion process more natural and flexible. Using this pipeline, we generated a dataset of 80k fusion data that encompass various scenarios, including placement, handheld, wearable, and style transfer tasks. Additionally, we introduce DreamFuse, an adaptive image fusion framework with the diffusion transformer. This framework incorporates positional affine transformations to encode the position and size of foreground, employs shared attention mechanisms to establish connections between the foreground and background, and ultimately leverages localized direct preference optimization to further enhance the quality of the fused images. Experimental results show that our method outperforms existing approaches on multiple benchmarks.

5. Acknowledgment

This work is supported in part by the National Key R&D Program of China (2024YFB3908503), and in part by the National Natural Science Foundation of China (62322608).

References

[1] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *NeurIPS*, 36:30286–30305, 2023. 2

[2] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, 2024. 2, 3, 6, 7

[3] Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. Zero-shot image editing with reference imitation. *NeurIPS*, 37:84010–84032, 2025. 6, 7

[4] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *CVPR*, pages 19790–19800, 2024. 4

[5] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *ICLR*, 2023. 2

[6] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *ICCV*, pages 7323–7334, 2023. 2

[7] Jixuan He, Wanhua Li, Ye Liu, Junsik Kim, Donglai Wei, and Hanspeter Pfister. Affordance-aware object insertion via mask-aware dual diffusion. *arXiv preprint arXiv:2412.14462*, 2024. 2, 3, 6, 7

[8] Zijian He, Peixin Chen, Guangrun Wang, Guanbin Li, Philip HS Torr, and Liang Lin. Wildvidfit: Video virtual try-on in the wild via image-based controlled diffusion models. In *ECCV*, pages 123–139. Springer, 2024. 2

[9] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 7

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 2, 4

[11] Junjia Huang, Pengxiang Yan, Jinhang Cai, Jiyang Liu, Zhao Wang, Yitong Wang, Xinglong Wu, and Guanbin Li. Dreamlayer: Simultaneous multi-layer generation via diffusion mode. *arXiv:2503.12838*, 2025. 4

[12] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024. 2

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 2

[14] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, pages 1931–1941, 2023. 2

[15] Black Forest Labs. Flux: Inference repository, 2024. Accessed: 2024-10-25. 3, 6

[16] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. *arXiv preprint arXiv:2404.04465*, 2024. 3

[17] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 5

[18] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025. 3, 5

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[20] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *ICCV*, pages 2294–2305, 2023. 2, 3, 6, 7

[21] Quanling Meng, Qinglin Liu, Zonglin Li, Xiangyuan Lan, Shengping Zhang, and Liqiang Nie. High-resolution image harmonization with adaptive-interval color transformation. In *NeurIPS*, 2024. 3

[22] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *CVPR*, pages 21033–21043, 2022. 2

[23] Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. *arXiv preprint arXiv:2306.06101*, 2023. 6

[24] Jinlong Peng, Zekun Luo, Liang Liu, and Boshen Zhang. Frih: fine-grained region-aware image harmonization. In *AAAI*, pages 4478–4486, 2024. 3

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[26] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 2

[29] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. 6

[30] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 3

[31] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, pages 2149–2159, 2022. 2

[32] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 3, 2024. 2, 3, 4

[33] Weijing Tao, Xiaofeng Yang, Miaomiao Cui, and Guosheng Lin. Motioncom: Automatic and motion-aware image composition with llm and video diffusion prior. *arXiv preprint arXiv:2409.10090*, 2024. 3

[34] Gemma Canet Tarrés, Zhe Lin, Zhifei Zhang, Jianming Zhang, Yizhi Song, Dan Ruta, Andrew Gilbert, John Collomosse, and Soo Ye Kim. Thinking outside the bbox: Unconstrained generative object compositing. In *ECCV*, 2024. 2

[35] Xueyun Tian, Wei Li, Bingbing Xu, Yige Yuan, Yuanzhuo Wang, and Huawei Shen. Mige: A unified framework for multimodal instruction-based image generation and editing. *arXiv preprint arXiv:2502.21291*, 2025. 3

[36] Petru-Daniel Tudosiu, Yongxin Yang, Shifeng Zhang, Fei Chen, Steven McDonagh, Gerasimos Lampouras, Ignacio Iacobacci, and Sarah Parisot. Mulan: A multi layer annotated dataset for controllable text-to-image generation. In *CVPR*, pages 22413–22422, 2024. 2

[37] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, pages 8228–8238, 2024. 3, 5

[38] Ke Wang, Michaël Gharbi, He Zhang, Zhihao Xia, and Eli Shechtman. Semi-supervised parametric real-world image harmonization. In *CVPR*, pages 5927–5936, 2023. 2

[39] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2

[40] Yibin Wang, Weizhong Zhang, Jianwei Zheng, and Cheng Jin. Primecomposer: Faster progressively combined diffusion for image composition with attention steering. In *ACM MM*, pages 10824–10832, 2024. 2, 3

[41] Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. In *ECCV*, pages 112–129. Springer, 2024.

[42] Daniel Winter, Asaf Shul, Matan Cohen, Dana Berman, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectmate: A recurrence prior for object insertion and subject-driven generation. *arXiv preprint arXiv:2412.08645*, 2024. 2, 3

[43] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *NeurIPS*, 36:15903–15935, 2023. 3, 7

[44] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024. 3, 7

[45] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, pages 18381–18391, 2023. 2

[46] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2

[47] Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. Controlcom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040*, 2023. 6, 7

[48] Jiacheng Zhang, Jie Wu, Huafeng Kuang, Haiming Zhang, Yuxi Ren, Weifeng Chen, Manlin Zhang, Xuefeng Xiao, and Guanbin Li. Treereward: Improve diffusion model via tree-structured feedback learning. In *ACM MM*, pages 4533–4542, 2024. 3

[49] Jiacheng Zhang, Jie Wu, Yuxi Ren, Xin Xia, Huafeng Kuang, Pan Xie, Jiashi Li, Xuefeng Xiao, Weilin Huang, Min Zheng, et al. Unifl: Improve stable diffusion via unified feedback learning. *NeurIPS*, 2025. 3

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *ICLR*, 2025. 2, 3

[51] Tao Zhang, Cheng Da, Kun Ding, Kun Jin, Yan Li, Tingting Gao, Di Zhang, Shiming Xiang, and Chunhong Pan. Diffusion model as a noise-aware latent reward model for step-level preference optimization. *arXiv preprint arXiv:2502.01051*, 2025. 3

[52] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *CVPR*, pages 8069–8078, 2024. 2

[53] Weizhi Zhong, Huan Yang, Zheng Liu, Huiguo He, Zijian He, Xuesong Niu, Di Zhang, and Guanbin Li. Mod-adapter: Tuning-free and versatile multi-concept personalization via modulation adapter. *arXiv:2505.18612*, 2025. 2

[54] Jing Zhou, Ziqi Yu, Zhongyun Bao, Gang Fu, Weilei He, Chao Liang, and Chunxia Xiao. Foreground harmonization and shadow generation for composite image. In *ACM MM*, pages 8267–8276, 2024. 3